Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures

Published in final edited form as:
Nat Genet. 2014 January ; 46(1): 82–87. doi:10.1038/ng.2848.

Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures

Tami D. Lieberman[1], **Kelly B. Flett**[2], **Idan Yelin**[3], **Thomas R. Martin**[4], **Alexander J. McAdam**[5], **Gregory P. Priebe**[2,6,7], and **Roy Kishony**[1,3]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[2]Division of Infectious Diseases, Department of Medicine, Boston Children's Hospital; and Harvard Medical School, Boston, MA 02115, USA

[3]Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel

[4]Division of Respiratory Diseases, Department of Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA

[5]Department of Laboratory Medicine, Boston Children's Hospital and Harvard Medical School, Boston, MA02115, USA

[6]Department of Anesthesiology, Perioperative and Pain Medicine, Division of Critical Care Medicine; Boston Children's Hospital; and Harvard Medical School, Boston, MA 02115, USA

[7]Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Abstract

Advances in sequencing have enabled the identification of mutations acquired by bacterial pathogens during infection[1-10]. However, it remains unclear whether adaptive mutations fix in the population or lead to pathogen diversification within the patient[11,12]. Here, we study the genotypic diversity of *Burkholderia dolosa* within people with cystic fibrosis by re-sequencing individual colonies and whole populations from single sputum samples. Extensive intra-sample diversity reveals that mutations rarely fix within a patient's pathogen population—instead, diversifying lineages coexist for many years. When strong selection is acting on a gene, multiple adaptive mutations arise but neither sweeps to fixation, generating lasting allele diversity that provides a recorded signature of past selection. Genes involved in outer-membrane components, iron scavenging and antibiotic resistance all showed this signature of within-patient selection. These results offer a general and rapid approach for identifying selective pressures acting on a pathogen in individual patients based on single clinical samples.

Two opposing models of within-patient bacterial evolution have been proposed: a “dominant lineage” model, in which beneficial mutations drive superior lineages to dominate in the population, and a “diverse community” model whereby adaptive lineages rise to intermediate frequency and coexist with other lineages (Fig. 1)[11-14]. The diversity of within-

Correspondence should be addressed to A.J.M. (alexander.mcadam@childrens.harvard.edu), G.P.P. (gregory.priebe@childrens.harvard.edu), or R.K. (roy_kishony@hms.harvard.edu).

Accession numbers: SRP030656

Author Contributions: TDL, AJM, GPP and RK designed the study. AJM and TRM collected clinical samples. KBF, TRM, AJM and GPP conducted chart review and provided medical information. TDL performed experiments. TDL, IY and RK wrote the sequence analysis scripts. TDL and RK analyzed the data. TDL, AJM, GPP and RK interpreted the results and wrote the paper.

URLs: *Burkholderia dolosa* sequencing project, http://www.broadinstitute.org/annotation/genome/burkholderia_dolosa

patient pathogen populations has major implications for drug treatment and resistance[7,15,16], for inferring transmission networks[8,9,17,18], and for understanding evolutionary processes[13,19]. Here, to distinguish between these models and to understand the sources of genetic diversity, we compared the genomes of many bacterial cells of the same strain from the same clinical sample.

We focused on chronic infections with *Burkholderia dolosa*, a rare and deadly opportunistic pathogen that spread among 39 people in with cystic fibrosis (CF) cared for at a single center in Boston starting in the 1990s[20,21]. The airways of these patients were infected with very similar starting strains, and surviving patients have been colonized for years. A previous retrospective study of single-colony isolates revealed specific *B. dolosa* genes that evolved under strong selective pressures during the outbreak[8]. Now, using sputum samples collected during clinical care, we characterize contemporary intraspecies diversity in 5 individuals from this outbreak who have been infected with *B. dolosa* since the early 2000's.

We used two genomic approaches, colony re-sequencing (Patient 1) and deep population sequencing (Patients 1-5), to identify single nucleotide mutations and their frequencies within single sputum samples. In our colony re-sequencing approach, we isolated dozens of colonies from a clinical sample and analyzed their genomes individually by alignment of reads to a *B. dolosa* reference genome, AU0158, a strain taken from a different patient in this outbreak. Since each colony originates from a single bacterium, this approach is equivalent to comparing different bacterial cells from the initial clinical sample. In the population sequencing approach[22,23], we pooled hundreds of colonies from each clinical sample and sequenced the pool with deep coverage (∼450×). We then aligned reads to AU0158 and identified fixed mutations, appearing in all reads, and polymorphisms, appearing in only a fraction of the reads. To remove false positive polymorphic sites caused by systematic sequencing or alignment errors[24,25], we developed a set of thresholds and statistical tests that reject polymorphic sites where the mutated and ancestral reads have significantly different properties[22,23] (see Supplementary Note). We calibrated this approach using an isogenic control for which we expect no polymorphisms. For validation, we performed both methods on a single sample from Patient 1, comparing diversity among 29 individual colonies to the population sequencing approach (Fig. 2). The population sequencing approach reliably detects polymorphisms where the minor allele frequency is larger than 3%, while decreasing the cost and labor required per sample.

We found that most mutations that arise during the course of infection do not fix, but remain polymorphic within the patient. The colony re-sequencing approach performed for Patient 1 revealed 188 mutations occurring in some, but not all, isolates and only 10 mutations shared among all isolates. This dominance of polymorphisms, also seen in the population sequencing from the same sample, strongly supports the diverse community model (Fig. 3a-b). Similarly, for the four other patients, population sequencing on single samples identified a preponderance of polymorphisms compared to fixed mutations (≥73% of mutations, Fig. 3b). We found these excesses despite the bias to overestimate mutations fixed during infection; some fixed mutations in a sputum sample might be polymorphic within the patient's airways or have fixed prior to patient colonization (Supplementary Fig. 1).

The observed genomic diversity is a reflection of multiple coexisting lineages. Investigating the community structure of *B. dolosa* within Patient 1, we found a deeply branched phylogeny with 6 lineages separated by at least 5 lineage-specific mutations (Fig. 3a). On average, pairs of isolates from this sample differed by 26 mutations, and, of all 406 possible isolate pairs, only one was identical. Thus, even within a single sputum sample, the population is so diverse that full identity between isolates is extremely rare.

In one patient (Patient 5), the *B. dolosa* community had many more mutations than other patients' populations ($P < 0.05$, Grubbs' test for outliers). This excess of mutations is due solely to increased transitions and not transversions, suggesting hypermutation (Supplementary Fig. 2a, $P < 0.01$, Grubbs' test). A search of the 199 mutated genes unique to Patient 5's population revealed a single mutation involved in DNA repair: a nonsynonymous mutation at a conserved position in *mutL*, defects of which are known to cause excess transitions[26] (Supplementary Fig. 2b). These excess mutations are enriched in synonymous mutations relative to the other patients, further supporting hypermutation ($P < .001$, Supplementary Fig. 2c). While hypermutation is a common phenotype in many pathogens, hypothesized to accelerate the evolution of antibiotic resistance[27-30], it has not been previously described in members of the *B. cepacia* complex[31].

For how long have these diverging lineages coexisted? The time to the last common ancestor (LCA) of each non-hypermutating patient's population[32] can be estimated using the number of mutations accumulated since the LCA and the molecular clock previously measured for this outbreak (2.1 SNPs/year[8]). Given the phylogeny of isolates from Patient 1, we calculated the distribution of the number of mutations since the LCA, d_{LCA}, across the population (Fig. 4a). The mean value of d_{LCA} across isolates, $\langle d_{LCA} \rangle$, is 19.6 single nucleotide mutations per genome (95% confidence interval, CI = 18.3-20.8), suggesting that the LCA existed 9.3 years ago (CI = 8.7-9.9). This places the LCA of the isolates from this sample slightly earlier than the first *B. dolosa* culture from this patient (7.6 years before sample collection), suggesting that the *B. dolosa* population in Patient 1 has been diverging since, or perhaps before, initial colonization. While the population sequencing approach cannot provide a distribution of d_{LCA}, due to a lack of information regarding linkage between mutations, we can still calculate $\langle d_{LCA} \rangle$: it is the sum of the polymorphic mutation frequencies (see Supplementary Note for derivation). Using this approach, the estimated time to LCA for Patient 1's population is 7.9 years. This value is slightly lower than calculated from the clonal re-sequencing approach, likely due to mutations left undetected by our conservative polymorphism caller (see Supplementary Note for discussion of error). For Patients 2 and 4, the time to LCA calculated by this population sequencing approach is several years less than the time since first positive culture, suggesting fixation events sometime during these patients' histories (Supplementary Table 1). For all these patients, we estimate that diverging lineages have coexisted in each of these patients for at least 5 years (Fig. 4b).

To explore the drivers of this long-coexisting diversity, we examined the identity of the evolving genes. Interestingly, we found that within each sample, several *B. dolosa* genes carried 2-4 coexisting polymorphisms (Supplementary Table 2). This clustering is a significant departure from a neutral model given the number of mutations and the distribution of gene lengths (Fig. 5a, $P < 0.005$ for Patients 1-4; **Online Methods**). A similar analysis at the operon-level further identified several operons enriched for polymorphisms (Supplementary Table 3 and Supplementary Fig. 3). An enrichment of nonsynonymous mutations in these multi-diverse genes and operons suggests that they are drivers of adaptive change *in vivo* (dN/dS = 7.0, CI = 2.3-34.9, Fig. 5b). Polymorphisms are thus concentrated within genes undergoing adaptive evolution.

To understand why polymorphisms cluster within some genes, we asked if coexisting mutations in the same gene appeared in different lineages or were linked in a double mutant. Examining the single isolate genomes, we found no isolates with doubly mutated genes (Supplementary Fig. 4). Similarly, for the population sequencing, in 10 of 11 cases where polymorphic positions are close enough on the genome to be covered by the same short sequencing reads, we did not find reads that contain both variants (Fig. 5c, Supplementary Fig. 5). In some of these cases, the ancestral genotype is completely purged from the

population (Fig. 5d). Thus, diversification is driven by multiple adaptive mutations in the same genes evolving in parallel within individual patients.

These findings provide a new signature of past selective pressures detectable in a single clinical sample; the coexistence of multiple polymorphisms within the same gene in a clinical sample. Sixteen *B. dolosa* genes display this multi-diverse signature, including genes with homologs involved in outer membrane synthesis, antibiotic resistance, iron scavenging, oxygen sensing, amino acid synthesis, lactate utilization, and stress response. Additionally, some genes with less characterized biological roles display a multi-diverse signature, including two transcriptional regulators with unknown targets in *B. dolosa*, an uncharacterized glucoamylase, and two genes that encode hypothetical proteins (Supplementary Table 2). A similar signature for selection is seen in three operons, two involved in lipopolysaccharide transport and one containing a two-component regulatory system with unknown targets (Supplementary Table 3). Selection on many of these elements can be rationalized based on the relevance of their annotated functions to conditions to which the bacteria are exposed in the course of the infection. Yet, further investigation will be required to understand the potential roles of some of these genes in antibiotic resistance, fitness, and other aspects of pathogenesis.

We found that many of the selective forces acting on the pathogen are the same across patients (Fig. 5e). Often, genes showing a multi-diverse signature for selection in one patient also carry mutations in other patients ($P < 0.002$, hypergeometric test). A prominent example is *gyrA*, a well-studied target of quinolones, which is mutated in all patient populations. Further support for commonality in mutational trajectories across patients emerges from a significant overlap between this list of 16 multi-diverse genes and 17 genes previously found to be under parallel evolution across a larger group of patients, only one of whom (Patient 2) was included in both studies ($P < 0.001$, hypergeometric test). Thus, the study of a single clinical sample can provide generalizable lists of selective pressures felt within the human body.

Yet, some multi-diverse signatures are patient-specific. A penicillin-binding protein (BDAG_01166, homologous to PBP7) has 3 nonsynonymous mutations in Patient 1, but is not mutated in other patients. Such patient-specific parallel evolution might reflect patient-specific selective pressure or perhaps a fitness benefit dependent upon previously acquired mutations. But these hypotheses are hard to test because the genomic target for a selective force might include more than one gene. For example, four of the five patients' populations have a mutation in a homolog of the histidine kinase *fixL* (BDAG_01161, known to be under strong selection in these infections[8]) while the fifth has a mutation in the corresponding response regulator.

To investigate the stability of these multi-diverse signatures for selection, we collected a second sputum sample 14 days after initial sample collection from Patient 2. Three of the four genes with the multi-diverse signature at day 0 show the same pattern at day 14. The absence of the signature in the fourth gene at the later time point does not reflect a relaxation in selection for mutant alleles, but rather incomplete detection of genes under selection; this gene also has abundant nonsynonymous mutants at day 14, concentrated at a single nucleotide position (Supplementary Fig. 6). These results suggest that the multi-diverse signature for selection is relatively stable and that multiple sample collections per patient can increase the sensitivity of the detection.

Our results reject the dominant lineage model of infection, yet demonstrate that these diversifying bacteria adapt under the pressure of natural selection. These observations are consistent with clonal interference: in large asexual populations, multiple beneficial

mutations emerge and compete, impeding the ability of these lineages to reach fixation[33-35]. In addition to large population size (10^8 cells/mL sputum), the branched structure of the airways may further hinder the capacity of any adaptive lineage to dominate and fix, and the immune system or niche-specific adaptations might directly promote diversity. Diversified by any of these means, lineages may then continue to evolve in parallel against common selective forces.

As *B. dolosa* adapts to the airways of people with cystic fibrosis, mutations lead to diversification rather than fixation and replacement. Though it is possible that adaptive mutations will lead to fixation more frequently in other infections, there is evidence that, at least in long-term colonization, diversity might be common[14,36-38]. This long-term coexistence of diverse lineages records the genomic history of selection on the pathogen within its host. The ability to rapidly read off within-patient evolutionary history from the genotypic diversity within a single clinical sample may greatly accelerate the ability to survey selective pressures acting on bacterial pathogens *in vivo* – shifting from an epidemic level investigation to a single-patient paradigm.

Online Methods

Study cohort and sample collection

An epidemic clone of *B. dolosa* infected and colonized 39 individuals with cystic fibrosis in the Boston area over a 20-year period[21]. We studied *B. dolosa* intrapatient diversity in 5 surviving individuals still infected with *B. dolosa*. All subjects were male, had homozygous ΔF508 mutations, had not received lung transplants, were between 21 and 35 years of age, and had been colonized for between 7 and 10 years at the time of sample collection (see Supplementary Table 1). Longitudinal microbial isolates from Patient 2 were also included in a previous retrospective study (patient J in reference 8).

For Patient 1, both the colony re-sequencing and deep population sequencing approaches were performed on a single sputum sample (P1). For Patient 2, population deep sequencing was performed on each of two sputum samples (P2 and P2T), collected 14 days apart. Between collections, Patient 2 was treated for a pulmonary exacerbation, including a change in antibiotic regimen, but his condition did not improve and *B. dolosa* density did not decrease. For Patients 3-5, population sequencing was performed on a single sputum sample from each patient (P3-P5).

Expectorated sputum samples were collected at Boston Children's Hospital after written informed consent was obtained under protocols approved by the Institutional Review Boards at Boston Children's Hospital and Harvard Medical School. Samples were liquefied with dithiothreitol[40] and stored at −80°C in 20% glycerol. *B. dolosa* was cultured from frozen samples. For population sequencing, a plate with 5,000 to 30,000 small colonies was chosen from a serial dilution. See Supplementary Note for more details on sample preparation.

Illumina sequencing

Genomic DNA was extracted using MoBio UltraClean Microbial DNA Isolation Kit per the manufacturer's instructions. Genomic libraries were constructed and barcoded using the Illumina-compatible Epicentre Nextera DNA Sample Prep Kit and following manufacturer's instructions (PCR amplification in the Nextera preparation does not introduce false positive polymorphisms, see Supplementary Note). Genomic libraries were sequenced on the Illumina HiSeq 2000 by Partners HealthCare Center for Personalized Genetic Medicine. Individual colonies were sequenced using single-end, 50 base-pair (bp) reads and pooled samples were sequenced using paired-end, 50bp reads. Reads were aligned to the *B. dolosa* draft genome AU0158 (GenBank accession number AAKY00000000, see URLs), belonging

to an isolate recovered from patient zero of the outbreak. AU0158 consists of 233 contigs on 3 scaffolds (*B. dolosa* has 3 chromosomes). Standard approaches were used for read filtering and alignment (Supplementary Note). See Supplementary Table 4 for coverage statistics.

Mutation identification, colony re-sequencing

An outgroup of 3 outbreak strains (A-0-0, G-9-8, and N-12-6d-$, previously sequenced[8]) was included in the analysis to identify mutations fixed among the 29 isolates from P1. We considered genomic positions at which at least one pair of isolates was discordant on the called base and both members of the pair had FQ scores less than −40 (FQ scores are produced by SAMtools[41]; lower values indicate agreement amongst reads). Genomic positions for which multiple isolates had multiple calls per isolate were discarded (likely duplication not represented in the reference). A best call was forced for each isolate (Supplementary Table 5) and the list of concatenated SNPs was inputted into the dnapars program in PHYLIP v3.69[42]. The resulting phylogeny was visualized the tree using Figtree (Fig. 3b).

Mutation identification, deep population sequencing

Fixed mutations within each patient's population were called using the same procedure as individual isolates, with a stricter quality score threshold (FQ < -282). Custom MATLAB scripts and SAMtools-produced pileup files were used to summarize all calls and their related quality scores at each genomic position (e.g. base quality, mapping quality, tail distance; see Supplementary Note). Using the isogenic control, multiple isolates from Patient 1, and an interactive MATLAB environment that enabled investigation of the raw data, we developed a set of filters to identify true-positive polymorphic positions with minor allele frequency above 3% (Supplementary Table 6). Thresholds were chosen to minimize false positives. See Supplementary Note and Supplementary Figs. 7-8 for description of filters and sensitivity analysis.

Estimation of $<d_{LCA}>$

For the colony-based approach, d_{LCA} was calculated for each isolate as the number of mutations received by that isolate normalized by the size of the callable genome. For this approach, the callable genome is the set of genomic positions with FQ score < -40. The confidence interval for $<d_{LCA}>$ presented for this approach is calculated according to a Poisson distribution. For the pool-based approach, $<d_{LCA}>$ was calculated as the sum of the mutation frequencies at each polymorphic position called within that population, normalized by the size of the callable genome (see Supplementary Note). For the pool-based approach, we define the callable genome as the set of positions that met the chosen thresholds for coverage, average base quality, average mapping quality, and average tail distance for each strand, irrespective of nucleotide call. See Supplementary Fig. 6b and the Supplementary Note for a discussion of sources of error in estimating $<d_{LCA}>$ and time to LCA.

Detection of parallel evolution within patients

We define genes with a multi-diverse signature of selection as genes for which within the same sputum sample there were multiple polymorphisms and multiple polymorphisms per 2000bp (to account for the fact that long genes are more likely to be mutated multiple times by chance). To determine whether the number of genes showing this signature was a significant departure from what expected in a neutral model, we performed for each sputum sample 1000 simulations in which we randomly shuffle the polymorphisms found across the callable genome, and calculate how many genes show the signature of selection (Fig. 5a).

This analysis was repeated at the operon and pathway levels, using the free version FgenesB to identify operons and subsystem annotations provided by SEED[43] as pathways (see Supplementary Figure 3). As in the gene analysis, we considered operons and pathways to have a signature for selection if they had both multiple polymorphisms and multiple polymorphisms per 2000 nucleotides with the same patient.

dN/dS

Mutations were classified as nonsynonymous (N) or synonymous (S) according to annotations provided in genbank file. For open reading frames in draft genome without a provided frame, we used BLAST and RefSeq to identify the most likely reading frame in the neighborhood of the found mutations. For each dNdS calculation, we used the particular spectrum of mutations observed to calculate the expected N/S (e.g. A->C mutations are 10.6 times more likely to cause an N than G->A mutations). The observed value of N/S was divided by this expectation to get dN/dS. Confidence intervals and p-values were calculated according to binomial sampling. The dNdS reported Fig. 5b groups together the mutations found in genes and operons under selection; the same calculation for only genes gives a dN/dS of 5.9 (95% CI = 1.9-29.6).

Parallel evolution across patients

We used the hypergeometric distribution to assess the significance of overlap between gene sets. Of 225 *B. dolosa* genes mutated in P1-P4, only 16 showed the multi-diverse signature for selection within patients and only 29 genes were mutated in multiple of these patients (fixed or polymorphic), yet 7 genes are in common between these lists (P=.0015) Similarly, 13 of these 225 genes were also found on a list of 17 genes evolved in parallel across patients in a previous study[8]. These 13 genes were enriched in the 16 genes under selection in in this study (5 gene overlap, P=.0009). When this analysis was repeated without mutations from P2 (Patient 2 also included in retrospective study), 11 of the 189 mutated genes were found in the previous study and 13 genes show a multi-diverse signature for selection. The overlap between these lists of 11 and 13 genes is smaller but still significant (4 genes; P=.0035).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Jean-Baptiste Michel and members of the Kishony lab for insightful discussions and support, to the team at Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) for Illumina sequencing, to Laura Williams and Adam Palmer for discussions and technical assistance, and to Ylaine Gerardin, Justin Meyer, Laura Stone, and Rebecca Ward for their comments on the manuscript. T.D.L. and G.P.P. were supported in part by grants from the Cystic Fibrosis Foundation (LIEBER12H0 to T.D.L and PRIEBE1310 to G.P.P). This work was funded in part by the US National of Institutes of Health (GM081617 to R.K.), the New England Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (NERCE; AI057159 to R.K.) and Hoffman-LaRoche.

References

1. Mwangi MM, et al. Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. Proc Natl Acad Sci U S A. 2007; 104:9451–6. [PubMed: 17517606]
2. Comas I, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2012; 44:106–10. [PubMed: 22179134]

3. Ford CB, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet. 2011; 43:482–6. [PubMed: 21516081]
4. Kennemann L, et al. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A. 2011; 108:5033–8. [PubMed: 21383187]
5. Young BC, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A. 2012; 109:4550–5. [PubMed: 22393007]
6. Huse HK, et al. Parallel evolution in Pseudomonas aeruginosa over 39,000 generations in vivo. MBio. 2010; 1
7. Snitkin ES, et al. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. Science Translational Medicine. 2012; 4:148ra116–148ra116.
8. Lieberman TD, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet. 2011; 43:1275–80. [PubMed: 22081229]
9. Didelot X, et al. Genomic evolution and transmission of Helicobacter pylori in two South African families. Proceedings of the National Academy of Sciences. 2013
10. Wilson DJ. Insights from genomics into bacterial pathogen populations. PLoS Pathog. 2012; 8:e1002874. [PubMed: 22969423]
11. Workentine M, Surette MG. Complex Pseudomonas Population Structure in Cystic Fibrosis Airway Infections. American journal of respiratory and critical care medicine. 2011; 183:1581–1583. [PubMed: 21693712]
12. Nguyen D, Singh PK. Evolving stealth: genetic adaptation of Pseudomonas aeruginosa during cystic fibrosis infections. Proc Natl Acad Sci U S A. 2006; 103:8305–6. [PubMed: 16717189]
13. Chung JC, et al. Genomic variation among contemporary Pseudomonas aeruginosa isolates from chronically infected cystic fibrosis patients. J Bacteriol. 2012; 194:4857–66. [PubMed: 22753054]
14. Workentine ML, et al. Phenotypic Heterogeneity of Pseudomonas aeruginosa Populations in a Cystic Fibrosis Patient. PloS one. 2013; 8:e60225. [PubMed: 23573242]
15. Foweraker JE, Laughton CR, Brown DF, Bilton D. Phenotypic variability of Pseudomonas aeruginosa in sputa from patients with acute infective exacerbation of cystic fibrosis and its impact on the validity of antimicrobial susceptibility testing. J Antimicrob Chemother. 2005; 55:921–7. [PubMed: 15883175]
16. Sun G, et al. Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. J Infect Dis. 2012; 206:1724–33. [PubMed: 22984115]
17. Harris SR, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. The Lancet infectious diseases. 2012
18. Walker TM, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet infectious diseases. 2012
19. Hansen SK, et al. Evolution and diversification of Pseudomonas aeruginosa in the paranasal sinuses of cystic fibrosis children have implications for chronic lung infection. The ISME journal. 2011
20. Vermis K, et al. Proposal to accommodate Burkholderia cepacia genomovar VI as Burkholderia dolosa sp. nov. International journal of systematic and evolutionary microbiology. 2004; 54:689–691. [PubMed: 15143009]
21. Kalish LA, et al. Impact of Burkholderia dolosa on lung function and survival in cystic fibrosis. Am J Respir Crit Care Med. 2006; 173:421–5. [PubMed: 16272450]
22. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013; 31:213–219.
23. Barrick JE, Lenski RE. Genome-wide mutational diversity in an evolving population of Escherichia coli. Cold Spring Harb Symp Quant Biol. 2009; 74:119–29. [PubMed: 19776167]
24. Pickrell JK, Gilad Y, Pritchard JK. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science. 2012; 335:1302. author reply 1302. [PubMed: 22422963]
25. Nakamura K, et al. Sequence-specific error profile of Illumina sequencers. Nucleic acids research. 2011; 39:e90–e90. [PubMed: 21576222]

26. Oliver A, Mena A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. Clin Microbiol Infect. 2010; 16:798–808. [PubMed: 20880409]
27. Oliver A, Canton R, Campo P, Baquero F, Blazquez J. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science. 2000; 288:1251–4. [PubMed: 10818002]
28. Jolivet-Gougeon A, et al. Bacterial hypermutation: clinical implications. J Med Microbiol. 2011; 60:563–73. [PubMed: 21349992]
29. Hoboth C, et al. Dynamics of adaptive microevolution of hypermutable Pseudomonas aeruginosa during chronic pulmonary infection in patients with cystic fibrosis. J Infect Dis. 2009; 200:118–30. [PubMed: 19459782]
30. Marvig RL, Johansen HK, Molin S, Jelsbak L. Genome Analysis of a Transmissible Lineage of Pseudomonas aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. PloS genetics. 2013; 9:e1003741. [PubMed: 24039595]
31. Pope CF, Gillespie SH, Moore JE, McHugh TD. Approaches to measure the fitness of Burkholderia cepacia complex isolates. J Med Microbiol. 2010; 59:679–86. [PubMed: 20185551]
32. Kingman JFC. On the genealogy of large populations. Journal of Applied Probability. 1982:27–43.
33. Fogle CA, Nagle JL, Desai MM. Clonal interference, multiple mutations and adaptation in large asexual populations. Genetics. 2008; 180:2163–73. [PubMed: 18832359]
34. Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998; 102-103:127–44. [PubMed: 9720276]
35. Hegreness M, Shoresh N, Hartl D, Kishony R. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science. 2006; 311:1615–7. [PubMed: 16543462]
36. Mowat E, et al. Pseudomonas aeruginosa population diversity and turnover in cystic fibrosis chronic infections. Am J Respir Crit Care Med. 2011; 183:1674–9. [PubMed: 21297072]
37. Schloissnig S, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013; 493:45–50. [PubMed: 23222524]
38. Ashish A, et al. Extensive diversification is a common feature of *Pseudomonas aeruginosa* populations during respiratory infections in cystic fibrosis. Journal of Cystic Fibrosis. 2013
39. Menard A, de Los Santos PE, Graindorge A, Cournoyer B. Architecture of Burkholderia cepacia complex sigma70 gene family: evidence of alternative primary and clade-specific factors, and genomic instability. BMC Genomics. 2007; 8:308. [PubMed: 17784948]
40. Guss AM, et al. Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. ISME J. 2011; 5:20–9. [PubMed: 20631810]
41. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
42. Felsenstein J. PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics. 1989; 5:164–166.
43. Aziz RK, et al. The RAST Server: rapid annotations using subsystems technology. BMC genomics. 2008; 9:75. [PubMed: 18261238]

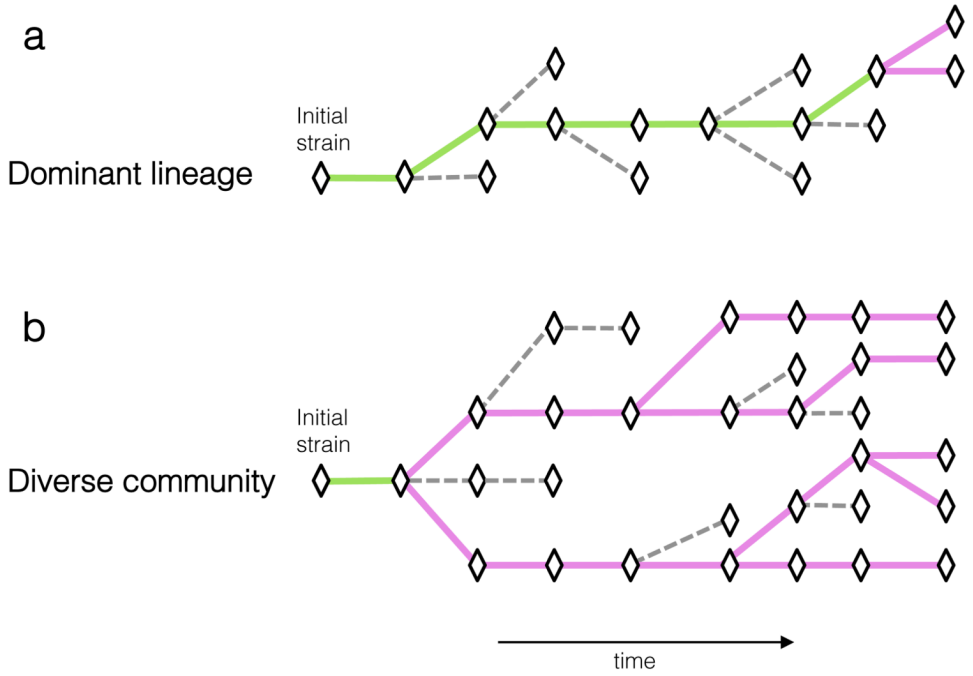

Figure 1. Alternative models of within-patient evolution

(**a**) In the dominant-lineage model of within-host evolution, lineages with beneficial mutations sweep to fixation (green lines), eliminating their less fit ancestors or other temporarily arising genotypes (dashed lines). In this model, most observed mutations will be fixed and polymorphic mutations will be rare, representing only recent mutational events (magenta lines). (**b**) In the diverse community-model, lineages coexist and compete for long stretches of time. In this model, most sampled mutations will be polymorphic.

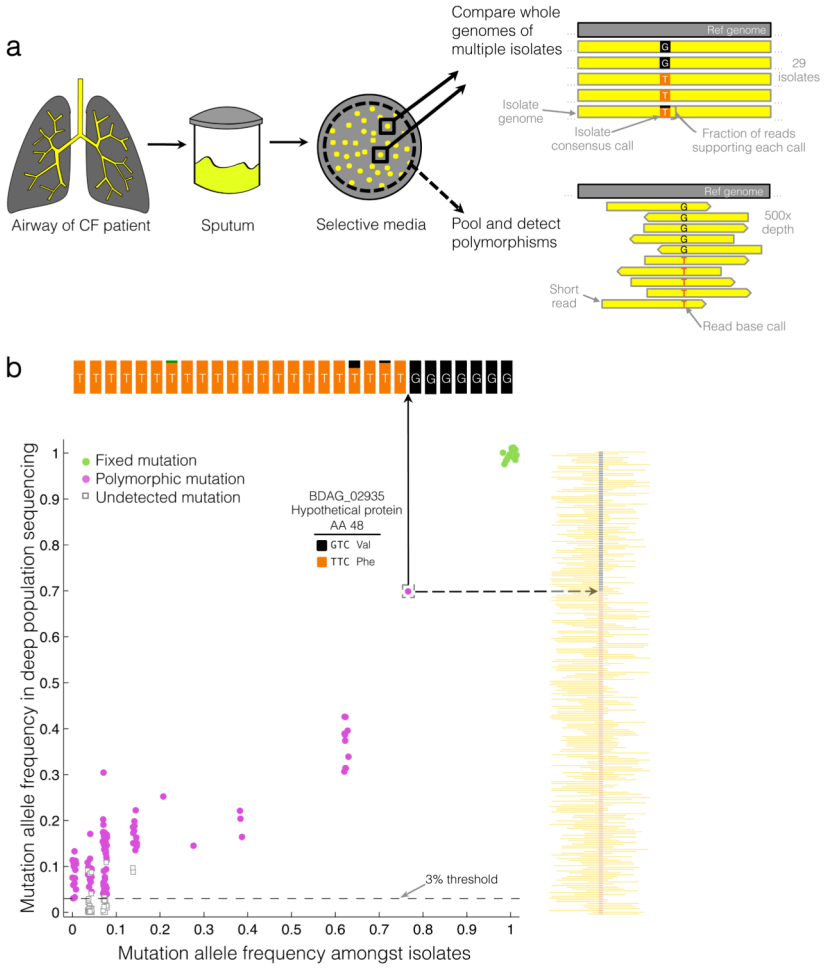

Figure 2. Two methods for studying genomic intraspecies diversity

(**a**) To study within-patient evolution, we cultured sputum samples from patients with cystic fibrosis on selective media. In the colony re-sequencing approach (solid arrows, performed for one patient), we isolated multiple individual colonies from the same single sample, independently called variants for each isolate via alignment of reads, and compared variants among the isolates. In the deep population sequencing approach (dashed arrow, performed for five patients), we pool hundreds of colonies from the same plate and analyze the pool's genomic DNA. We identified positions on the genome where some reads, originating from different colonies on the plate, disagree with an inferred ancestral genome (Online Methods). (**b**) Allele frequency estimates in the population sequencing (y-axis) versus the colony re-sequencing (x-axis) from the same sputum sample (P1) for each mutated position. Mutations are classified as either fixed (green circles) or polymorphic (magenta circles). Some mutations found in the colony-based approach are sub-threshold in frequency or confidence in the pool-based approach (open squares). Slight jitter is added in the X and Y locations for each point to improve visibility (up to 2% change). As an example, the insets at top and at right display a summary of the raw data at the indicated genomic position. The population sequencing (right) at this position shows 70% aligned reads supporting a T (orange) and 30% supporting a G (black), consistent with the corresponding number of colonies in the individual isolates (22, T; 7, G). Reads from each isolate (top) are mostly of identical calls (all T, or all G). Green indicates a single read in one isolate supporting an A,

likely a sequencing error. For further comparison of the two methods, see Supplementary Figure 7.

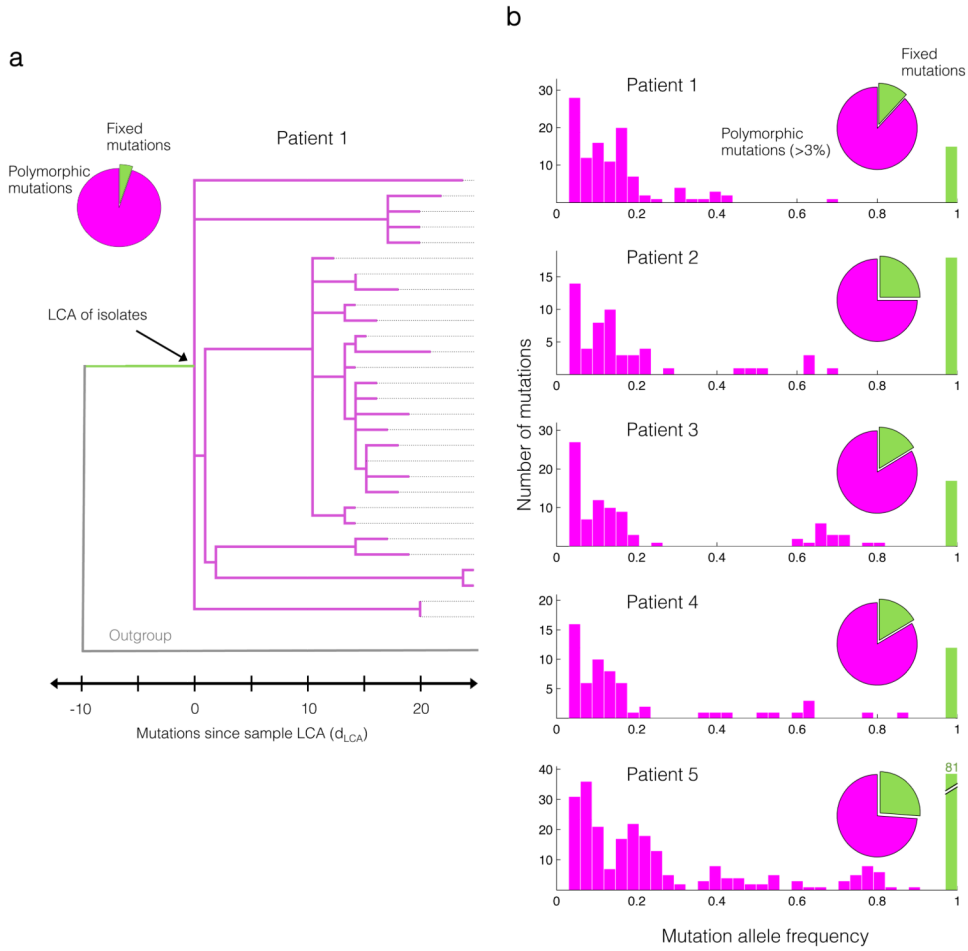

Figure 3. Within-patient evolution leads to diversification, not substitution

Mutations found in *B. dolosa* within-patient populations relative to the outgroup are classified as fixed (green), or polymorphic (magenta). An excess of polymorphic versus fixed mutations supports the diverse-community model over the dominant-lineage model. (**a**) A maximum-parsimony phylogeny of 29 isolates from the same sputum sample (P1) shows the coexistence of diverse sub-lineages separated by many single nucleotide mutations accumulating since the last common ancestor (LCA) of this patient. Each isolate is represented by a dotted line. (**b**) The diverse-community model is also supported by the distribution of allele frequencies from the population sequencing in 5 patients' samples.

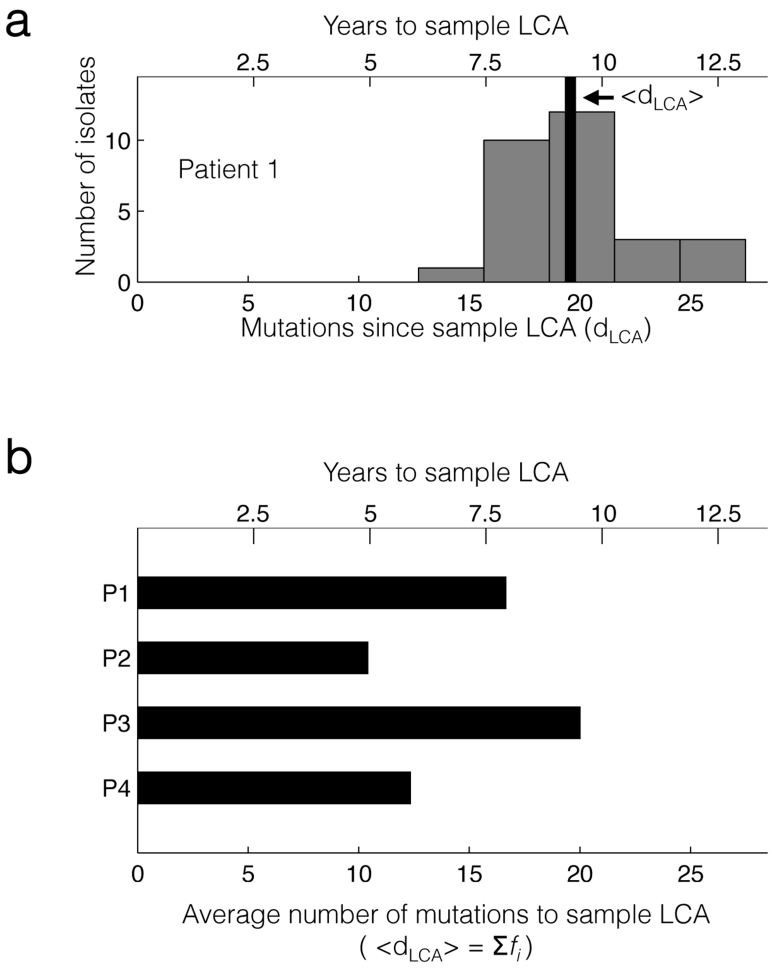

Figure 4. Sublineages coexist within a patient for many years after divergence

(**a**) A histogram of the number (d_{LCA}) of single nucleotide mutations found in isolates from Patient 1, relative to their LCA. The black bar indicates the mean value of d_{LCA} across the isolates. (**b**) The value of $<d_{LCA}>$ from the population sequencing data for patients Patients 1 through 4 (Online Methods). In both panels, the axis at top shows the relationship between d_{LCA} and years to LCA, as calculated via the molecular clock (2.1 SNPs/yr)[8].

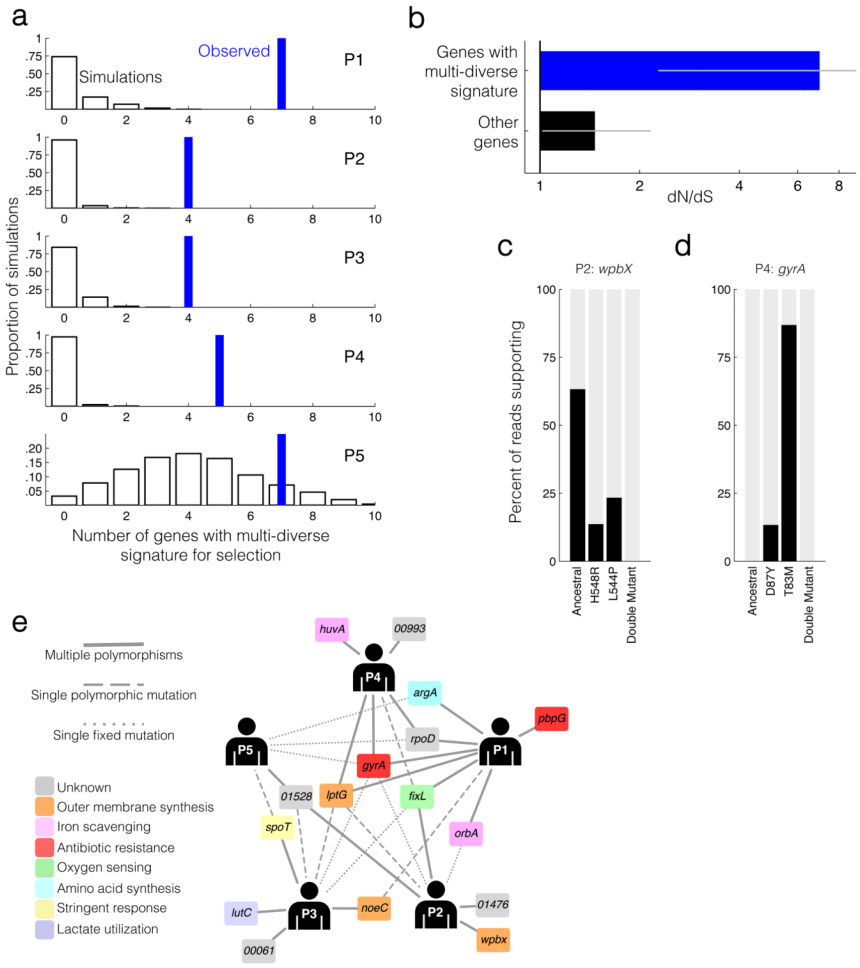

Figure 5. Coexistence of alternative adaptive mutations in the same sample highlights specific genes as drivers of within-host evolution

(**a**) Number of multi-diverse genes observed in samples from Patients 1-5 (P1-P5, blue bars) relative to a null expectation in which diverse sites are randomly distributed across the genome (histogram, 1000 simulations). For P5, the number of multi-diverse genes observed is not significant. (**b**) The canonical signal for selection, dN/dS, across the set of 16 genes and 3 operons showing a multi-diverse signature in at least one patient (P1-P4, 21 genes total, blue) versus dNdS across the set of genes not showing this signature (black). dN/dS >1 indicates positive selection for amino acid change. Error bars indicate 95% CIs. See Online Methods for details on the calculation of dNdS. (**c-d**) Linkage between nearby polymorphisms based on jointly overlapping short reads. Percentages of reads supporting the ancestral genotype, each of the single mutants, and the double mutant are plotted. No reads supporting the double mutant were found (c, n=524; d, n=415; See Supplementary Fig. 5 for exception). (e) A network of patients and genes showing a multi-diverse signature at least once in P1-P4. A gene is connected to a patient if it was mutated multiple times (solid line), had a single polymorphic mutation (dashed lined), or single fixed mutation (dotted line) within that patient. Genes closer to the center of the network are mutated in more patients, representing common targets of *in vivo* pathogen selection, while genes connected to single patients may indicate patient-specific adaptation. Genes are labeled with their closest homolog and predicted biological role. The biological role of r*poD* is unclassified because it is recently duplicated in the *B. cepacia* complex[39] (see Supplementary Note, Supplementary Table 2).